16
Changes in Utility as Information

In previous lectures we encountered many situations in which a decision maker, before making a decision, has the opportunity to observe data X, related to the state of nature of interest θ. The observation of X can give some information about the state of nature, and therefore be useful in decision making. In this lecture, we discuss how to quantify the value of this information. Our discussion is based on DeGroot [1].

16.1 Decision trees for statistical analyses

As we pointed out in Lecture 2, from a statistical viewpoint, interesting questions arise when there is the possibility of performing an experiment $\mathcal{E}$. Say $\mathcal{E}$ consists of observing a random variable X, with possible values in the *sample space* $\mathcal{X}$, and whose probability density function (or probability mass) is $p(x \mid \theta)$. As usual, θ is a *parameter* with possible values in a *parameter space* Θ.

Decision problems where an observation X, related to the parameter θ through the likelihood function, is available, are called *statistical decision problems*. Decision trees can represent a statistical decision problem, as depicted in Figure 16.1.1. Therefore, we can solve a statistical decision problem with dynamic programming, as we described in Lecture 14.

Solving the decision tree will tell us whether the experiment is worthwhile. Suppose the experiment has a cost c. The largest value of c such

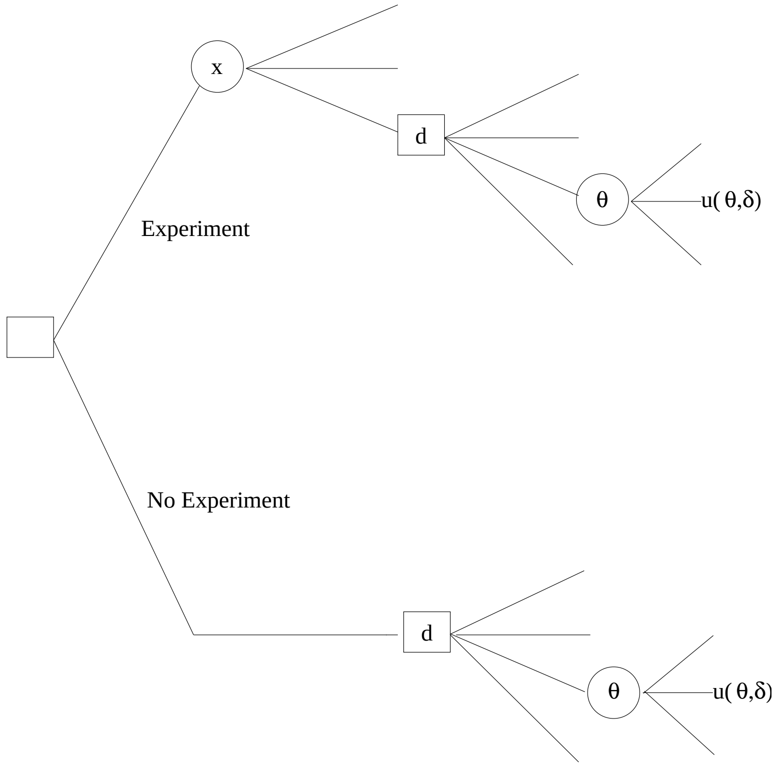

FIGURE 16.1.1. Decision tree for a statistical design problem

that the experiment is worthwhile can be thought of as the value of the experiment in this decision problem.

From this viewpoint, an experiment has value only if it may affect subsequent decision. This value can be measured based on the utilities assigned to the final outcome of the decision process. We now make this more formal.

16.2 Information

Let us start by considering the "No Experiment" branch of the tree. We know that we will choose an action a which has the maximum expected utility

$$U(\pi, a) = \int u(\theta, a)\pi(\theta)d\theta.$$

Let us call $V(\pi)$ the maximum value of the expected utility, under prior π for θ, that is:

$$V(\pi) = \sup_a U(\pi, a) \tag{16.2.1}$$

This is the amount of utility we expect to hold at the end of the problem.

Example 16.2.1 Let us illustrate $V(\pi)$ using specific utility functions.

a) $u(\theta, a) = -(\theta - a)^2$. As discussed in Lecture 2, the optimal decision is $a^* = E(\theta)$. Therefore,

$$\begin{aligned} V(\pi) &= -\int (\theta - a^*)^2 \pi(\theta) d\theta \\ &= -\int (\theta - E(\theta))^2 \pi(\theta) d\theta \\ &= -Var(\theta) \end{aligned}$$

b) $u(\theta, a) = \log(a(\theta))$, and $a(\theta)$ represents a distribution on θ.
As discussed in Lecture 4, this utility is a proper local scoring rule and so, the optimal decision is $a^*(\theta) = \pi(\theta)$.
Then,

$$V(\pi) = \int log(\pi(\theta)) \pi(\theta) d\theta$$

The negative of this quantity is also known as the *entropy* of the density $\pi(\theta)$.

c) $u(\theta, a) = I_{a=\theta}$, where $\Theta = \{\theta_1, \theta_2, \cdots\}$.
We have: $\max_a u(\pi, a) = \max_a \pi(a)$. Therefore, $V(\pi) = mode(\pi)$.

Theorem 16.2.1 *The function $V(\pi)$ is convex, that is, for any two distributions π_1 and π_2 of θ and $0 < \alpha < 1$:*

$$V(\alpha\pi_1 + (1-\alpha)\pi_2) \leq \alpha V(\pi_1) + (1-\alpha) V(\pi_2) \tag{16.2.2}$$

Proof: [1]

$$\begin{aligned} &V(\alpha\pi_1 + (1-\alpha)\pi_2) \\ &= \sup_{a \in \mathcal{A}} \int u(\theta, a) \left[\alpha\pi_1(\theta) + (1-\alpha)\pi_2(\theta)\right] d\theta \\ &= \sup_{a \in \mathcal{A}} \left[\alpha \int u(\theta, a)\pi_1(\theta) d\theta + (1-\alpha) \int u(\theta, a)\pi_2(\theta) d\theta\right] \\ &\leq \alpha \left[\sup_{a \in \mathcal{A}} \int u(\theta, a)\pi_1(\theta) d\theta\right] + (1-\alpha) \left[\sup_{a \in \mathcal{A}} \int u(\theta, a)\pi_2(\theta) d\theta\right] \\ &= \alpha V(\pi_1) + (1-\alpha) V(\pi_2) \end{aligned}$$

In addition, when we have the option to perform an experiment two questions can be considered:

1) After observing x, how much did we learn about θ?

2) How much do we expect to learn from X prior to observing it?

To simplify the notation, let us denote by π_x the posterior probability density function (or posterior probability mass in the discrete case) of θ, that is, $\pi_x(\theta) = \pi(\theta \mid x)$.

We can answer question 1 by evaluating the *observed information* about θ provided by the observation x.

We could define the observed information as the observed change in utility, that is,

$$V(\pi_x) - V(\pi). \tag{16.2.3}$$

Since it is possible that the posterior distribution of θ leaves the decision maker with more uncertainty about θ, with definition 16.2.2, the observed information can be either positive or negative. Alternatively, one can use:

$$\mathcal{I}(\pi \to \pi_x) = V(\pi_x) - u(\pi_x, a^*) \tag{16.2.4}$$

where a^* represents the Bayes decision with respect to the prior distribution π. In words, this definition tells us that the observed information is the expected difference, calculated with respect to the posterior distribution of θ, between the utility of the Bayes decision and the utility of the decision a^* that would be chosen if observation x had not been available. It turns out that with definition 16.2.3, the observed information is always non-negative.

Let us explore the notion of information in the following example.

Example 16.2.2 [Continued]

a) We have seen that in an estimation problem with utility $u(\theta, a) = -(\theta - a)^2$, $V(\pi) = -Var_\pi(\theta)$. The observed information (using definition 16.2.3) is:

$$\begin{aligned}
\mathcal{I}(\pi \to \pi_x) &= V(\pi_x) - u(\pi_x, a^*) \\
&= -Var_{\pi_x}(\theta) - u(\pi_x, E(\theta)) \\
&= -Var_{\pi_x}(\theta) - \left[-\int (\theta - E(\theta))^2 \pi_x(\theta) d\theta \right] \\
&= -Var_{\pi_x}(\theta) + [Var_{\pi_x}(\theta) + (E(\theta \mid x) - E(\theta))^2] \\
&= (E(\theta \mid x) - E(\theta))^2
\end{aligned}$$

b) When $u(\theta, a) = \log(a(\theta))$, we saw that $V(\pi) = \int \log(\pi(\theta))\pi(\theta) d\theta$. So,

$$\begin{aligned}
\mathcal{I}(\pi \to \pi_x) &= V(\pi_x) - u(\pi_x, a^*) \\
&= \int \log(\pi_x(\theta))\pi_x(\theta) d\theta - \int \log(\pi(\theta))\pi(\theta) d\theta \\
&= \int \pi_x(\theta) \log\left(\frac{\pi_x(\theta)}{\pi(\theta)} \right) d\theta
\end{aligned}$$

This expression is known as the Kullback-Leibler divergence between π_x and π. We will develop this case further in the next lecture.

c) If $u(\theta, a) = I_{a=\theta}$, as we saw, $V(\pi) = mode(\pi)$. Let us denote by θ^0 and θ^1 the modes for π and π_x, respectively. Then

$$\begin{aligned} \mathcal{I}(\pi \to \pi_x) &= V(\pi_x) - u(\pi_x, a^*) \\ &= \pi_x(\theta^1) - u(\pi_x, \theta^0) \\ &= \pi_x(\theta^1) - \pi_x(\theta^0) \end{aligned}$$

A natural way to answer question 2 is to compute the expected information, calculated with respect to the marginal distribution of X. In symbols, the expected information is:

$$E(\mathcal{I}(\pi \to \pi_x))$$

Taking the expectation of 16.2.3, we have

$$E(\mathcal{I}(\pi \to \pi_x)) = E(V(\pi_x)) - V(\pi) \tag{16.2.5}$$

Working out the expectation in the second term of the right hand side, we have:

$$\begin{aligned} E(u(\pi_x, a^*)) &= \int u(\pi_x, a^*) p(x) dx \\ &= \int \int u(\theta, a^*) \pi_x(\theta) p(x) d\theta dx \\ &= \int \int u(\theta, a^*) \frac{p(x \mid \theta)\pi(\theta)}{p(x)} p(x) d\theta dx \\ &= \int u(\theta, a^*) \pi(\theta) \int p(x \mid \theta) dx d\theta \\ &= \int u(\theta, a^*) \pi(\theta) d\theta \\ &= u(\pi, a^*) \\ &= V(\pi) \end{aligned} \tag{16.2.6}$$

leading to the desired result.

Consider the situation where you first observe X, and after that, you observe Y. The following theorem states that the expected information of the two experiments is additive.

Theorem 16.2.2

$$E(\mathcal{I}(\pi \to \pi_{xy})) = E(\mathcal{I}(\pi \to \pi_x)) + E(\mathcal{I}(\pi_x \to \pi_{xy}))$$

Proof: Let us denote by a_0 and a_1 the Bayes' decisions under π and π_x, respectively.

Using the definition of observed information 16.2.3, we have

$$\begin{aligned}
\mathcal{I}(\pi \to \pi_x) &= V(\pi_x) - u(\pi_x, a_0) \\
\mathcal{I}(\pi_x \to \pi_{xy}) &= V(\pi_{xy}) - u(\pi_{xy}, a_1) \\
\mathcal{I}(\pi \to \pi_{xy}) &= V(\pi_{xy}) - u(\pi_{xy}, a_0)
\end{aligned}$$

Observe that, by applying 16.2.5,

$$\begin{aligned}
E_x(u(\pi_x, a_0)) &= V(\pi) \\
E_y(u(\pi_{xy}, a_1)) &= V(\pi_x), \text{ for fixed x} \\
E_{xy}(u(\pi_{xy}, a_0)) &= V(\pi)
\end{aligned}$$

Now, if we compute the expected information (that is, take the expectations with respect to X and Y in the observed information), we obtain:

$$\begin{aligned}
E(\mathcal{I}(\pi \to \pi_{xy})) &= E(V(\pi_{xy})) - E(u(\pi_{xy}, a^*)) \\
&= E(V(\pi_{xy})) + V(\pi) \\
&= [E(V(\pi_{xy}) - E(V(\pi_x))] + [E(V(\pi_x)) - V(\pi)] \\
&= E(\mathcal{I}(\pi_x \to \pi_{xy})) + E(\mathcal{I}(\pi \to \pi_x))
\end{aligned}$$

In general, even if X and Y are iid given θ, it will not be true that the two terms are equal. We will explore this theme further in Lecture 17.

16.3 Distribution of Information

The study of the distribution of the information can be useful for model validation and design of experiments. We will consider an example based on the normal distributions.

Example 16.3.1 Suppose that a random quantity X is drawn from a $N(\theta, 1/r)$ distribution and that the prior distribution for θ is a $N(\mu_0, \tau_0)$. The posterior distribution of θ given $X = x$ is a $N(\mu_1, \sigma_1^2)$ where:

$$\mu_1 = \frac{\tau_0 \mu_0 + rx}{\tau_0 + r}$$

noindent and

$$\sigma_1^2 = \frac{1}{\tau_0 + r}$$

.

The marginal distribution of X is a $N(\mu_0, \frac{\tau_0 + r}{\tau r})$.

In example 16.2.2 we saw that with the utility function $u(\theta, a) = -(\theta - a)^2$, we have: $\mathcal{I}(\pi \to \pi_x) = (E(\theta \mid x) - E(\theta))^2$.

Therefore, for this problem, $\mathcal{I}(\pi \to \pi_x) = (\mu_1 - \mu_0)^2$. Let us study the distribution of $(\mu_1 - \mu_0)^2$.

	H_0	H_1
$a_0 =$ accept H_0	0	1
$a_1 =$ accept H_1	3	0

TABLE 16.4.1. Loss table for the hypotheses testing problem

Observe that:

$$\begin{aligned}\mu_1 - \mu_0 =& \frac{\tau_0\mu_0 + rx}{\tau_0 + r} - \mu_0 \\ =& \frac{r}{\tau_0}(x - \mu_0)\end{aligned}$$

Since $(x - \mu_0) \sim N(0, \frac{\tau_0 + r}{\tau_0 r})$, we have: $\mu_1 - \mu_0 \sim N(0, \frac{r}{\tau_0(\tau_0 + r)})$. Therefore,

$$\frac{\tau_0(\tau_0 + r)}{r}(\mu_1 - \mu_0)^2 \sim \chi_1^2$$

This let us make predictions about the amount of information contained in the experiment X.

16.4 Worked Exercise

Example 16.4.1 *$x_1, \cdots, x_9$ are a sample from a $N(\theta, 1)$. The prior π_0 is a $N(0, 4)$. You are interested in testing: $H_0 : \theta < 0$, versus $H_1 : \theta \geq 0$ based on the loss table shown in Table 16.4.1.*

Let $\Theta = \{H_0, H_1\}$ and $\mathcal{A} = \{a_0, a_1\}$. We have: $\pi(H_0) = \pi(\theta < 0) = 1/2$ and $\pi(H_1) = \pi(\theta \geq 0) = 1/2$.

Since the $x_1, \cdots, x_9$ are a sample from a $N(\theta, 1)$ and the prior π_0 is a $N(0, 4)$, we have $\pi(\theta \mid x_1, \cdots, x_n) \sim N(\frac{36}{37}\bar{x}, \frac{4}{37})$. By using this result we obtain,

$$\begin{aligned}\pi(H_0 \mid x) &= \pi(\theta < 0 \mid x) = \Phi\left(\frac{-18\bar{x}}{\sqrt{37}}\right) \\ \pi(H_1 \mid x) &= \pi(\theta \geq 0 \mid x) = 1 - \Phi\left(\frac{-18\bar{x}}{\sqrt{37}}\right)\end{aligned}$$

where $\Phi(x) = P(Z \leq x)$ and $Z \sim N(0, 1)$.

The expected utilities for each decision are:

- *for decision a_0:*

$$\begin{aligned}u(\pi, a_0) &= u(H_0, a_0)\pi(H_0) + u(H_1, a_0)\pi(H_1) \\ &= 0\pi(H_0) - 1\pi(H_1) = -1/2\end{aligned}$$

- *for decision a_1:*

$$\begin{aligned}u(\pi, a_1) &= u(H_0, a_1)\pi(H_0) + u(H_1, a_1)\pi(H_1) \\ &= -3\pi(H_0) + 0\pi(H_1) = -3/2\end{aligned}$$

Therefore, $V(\pi) = \sup_{a \in \mathcal{A}} u(\pi, a) = -1/2$.

If no observation is available, then $a^* = a_0$*, because it maximizes the expected utility.*

After observing x*, the expected utilities of the two decisions are:*

- *for decision* a_0*:*

$$\begin{aligned} u(\pi_x, a_0) &= u(H_0, a_0)\pi_x(H_0) + u(H_1, a_0)\pi_x(H_1) \\ &= 0\pi_x(H_0) - 1\pi_x(H_1) \\ &= \Phi\left(\frac{-18\bar{x}}{\sqrt{37}}\right) - 1 \end{aligned}$$

- *for decision* a_1*:*

$$\begin{aligned} u(\pi_x, a_1) &= u(H_0, a_1)\pi_x(H_0) + u(H_1, a_1)\pi_x(H_1) \\ &= -3\pi_x(H_0) + 0\pi_x(H_1) \\ &= -3\Phi\left(\frac{-18\bar{x}}{\sqrt{37}}\right) \end{aligned}$$

We will choose a_0 *if* $u(\pi_x, a_0) > u(\pi_x, a_1)$*, that is, if* $\Phi\left(\frac{-18\bar{x}}{\sqrt{37}})\right) - 1 > -3\Phi\left(\frac{-18\bar{x}}{\sqrt{37}}\right)$*. Rearranging, we get* $u(\pi_x, a_0) > u(\pi_x, a_1) \Leftrightarrow \bar{x} < 0.2279$

Since: $V(\pi_x) = \max\{u(\pi_x, a_0), u(\pi_x, a_1)\}$*, we have:*

$$V(\pi_x) = \begin{cases} u(\pi_x, a_0) = \Phi\left(\frac{-18\bar{x}}{\sqrt{37}}\right) - 1, & \text{if } \bar{x} < 0.2279 \\ u(\pi_x, a_1) = -3\Phi\left(\frac{-18\bar{x}}{\sqrt{37}}\right), & \text{if } \bar{x} \geq 0.2279 \end{cases}$$

Observe now that:

$$\begin{aligned} u(\pi_x, a^*) &= u(\pi_x, a_0) = p(H_0 \mid x)u(H_0, a_0) + p(H_1 \mid x)u(H_1, a_0) \\ &= p(H_0 \mid x)0 + p(H_1 \mid x)(-1) = -p(H_1 \mid x) \\ &= -1 + \Phi\left(\frac{-18\bar{x}}{\sqrt{37}}\right). \end{aligned}$$

Therefore:

$$\mathcal{I}(\pi \to \pi_x) = V(\pi_x) - u(\pi_x, a_0) = \begin{cases} 0, & \text{if } \bar{x} < 0.2279 \\ 1 - 4\Phi(\frac{-18\bar{x}}{\sqrt{37}})), & \text{if } \bar{x} \geq 0.2279 \end{cases}$$

The resulting observed information is shown in Figure 16.4.1.

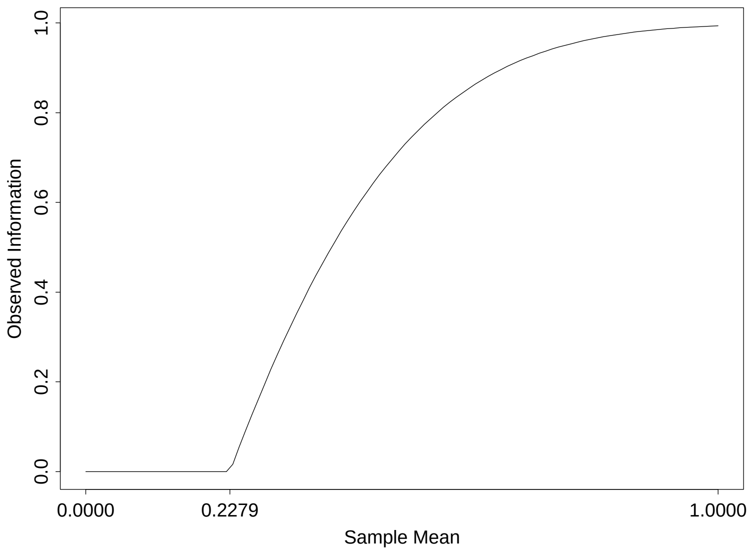

FIGURE 16.4.1. Observed Information

References

[1] M. H. DeGroot. Changes in utility as information. *Theory and Decision*, 17:287–303, 1984.